Language Models for MT
Original vs. Translated Text

G. Lembersky, N. Ordan, S. Wintner

Translation studies tell us...

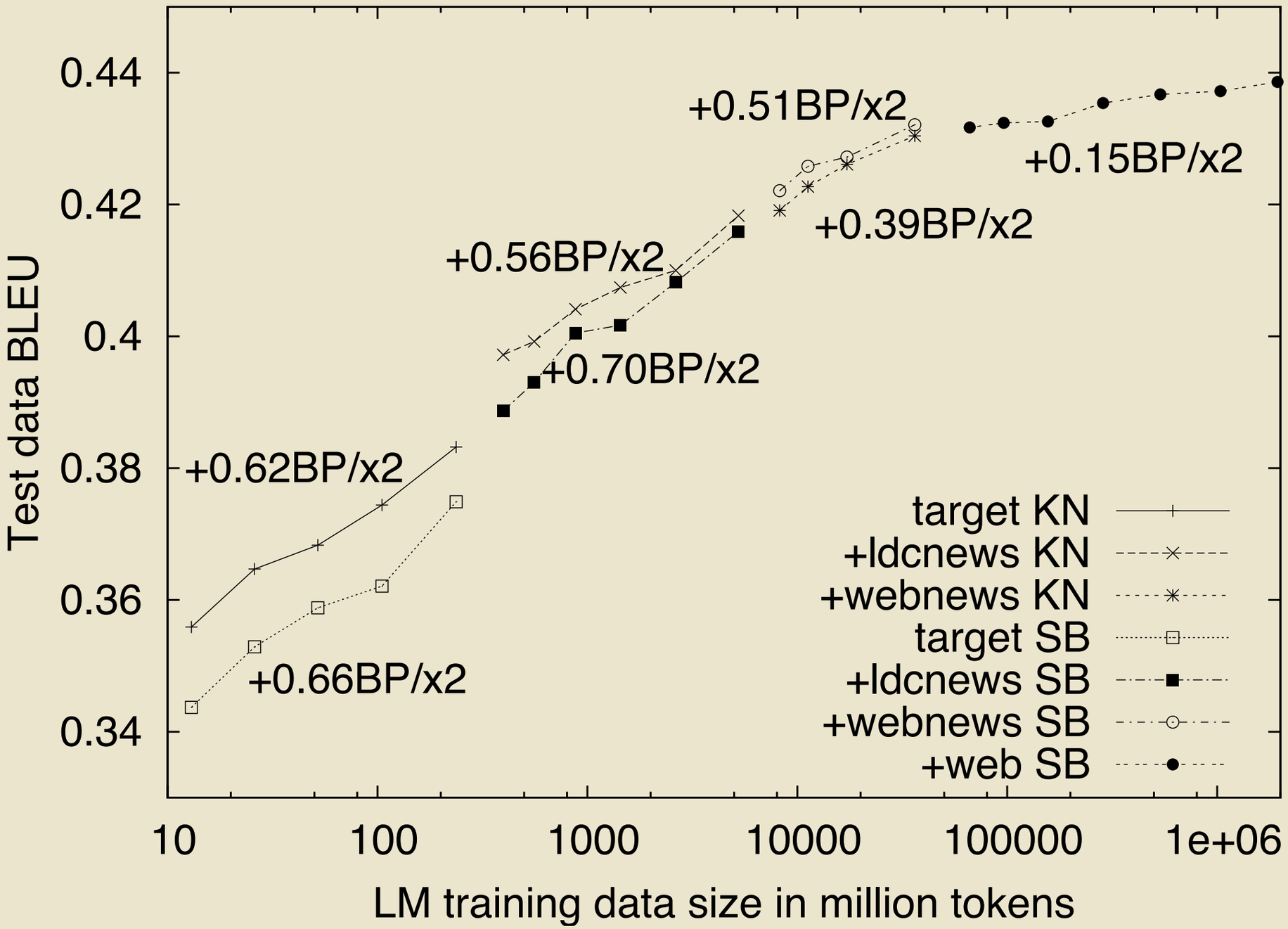
Translated texts ≠ original texts

- interference TM
- standardization LM

Translations share universals (translationalese)

- simplification
- explicitation

More data = better data?



Hypotheses

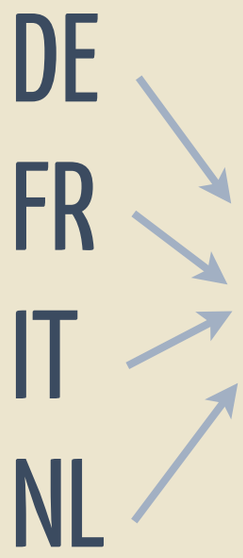
1. $S \rightarrow T \neq 0$

2. $S_1 \rightarrow T_1$
   $S_2 \rightarrow T_2$
   $(T_1 \sim T_2) \nsim 0$

3. $T > 0$ for $S \rightarrow T$

Corpora

1. Europarl



2.5 M

2. Hansards

FR → EN

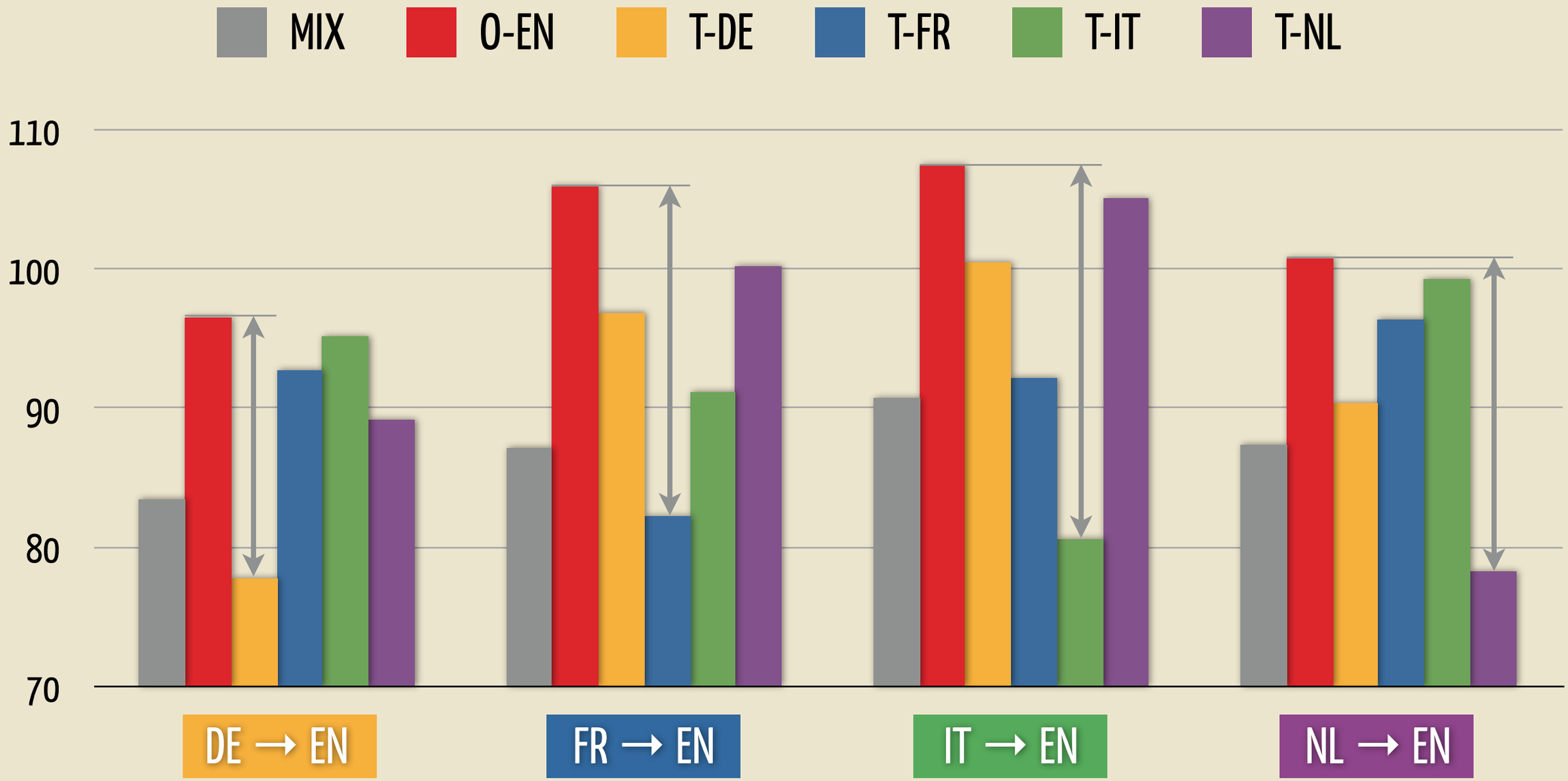
10 M

3. Hebrew-English

HE → EN

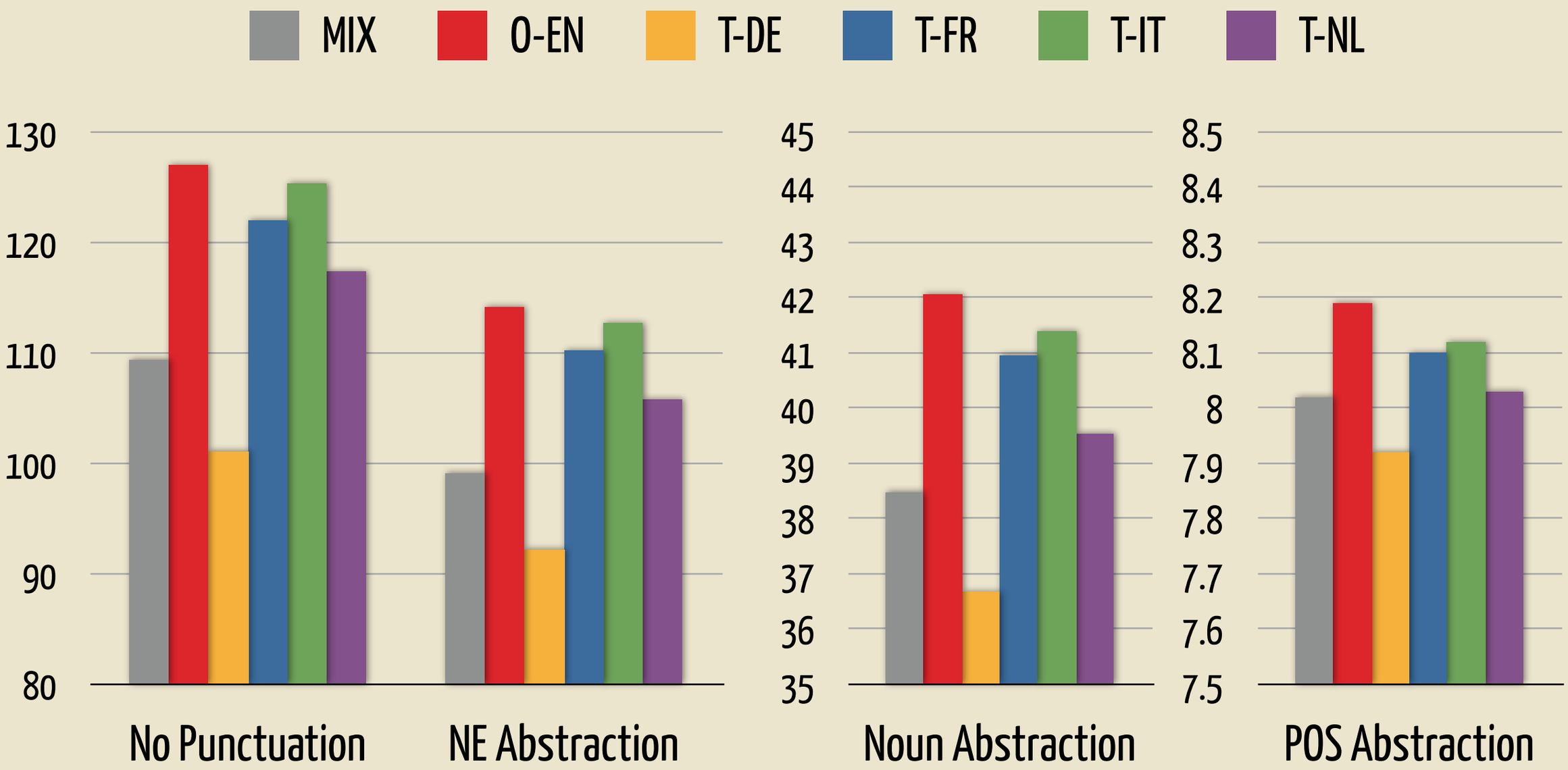
3.5 M

1. Perplexity experiments

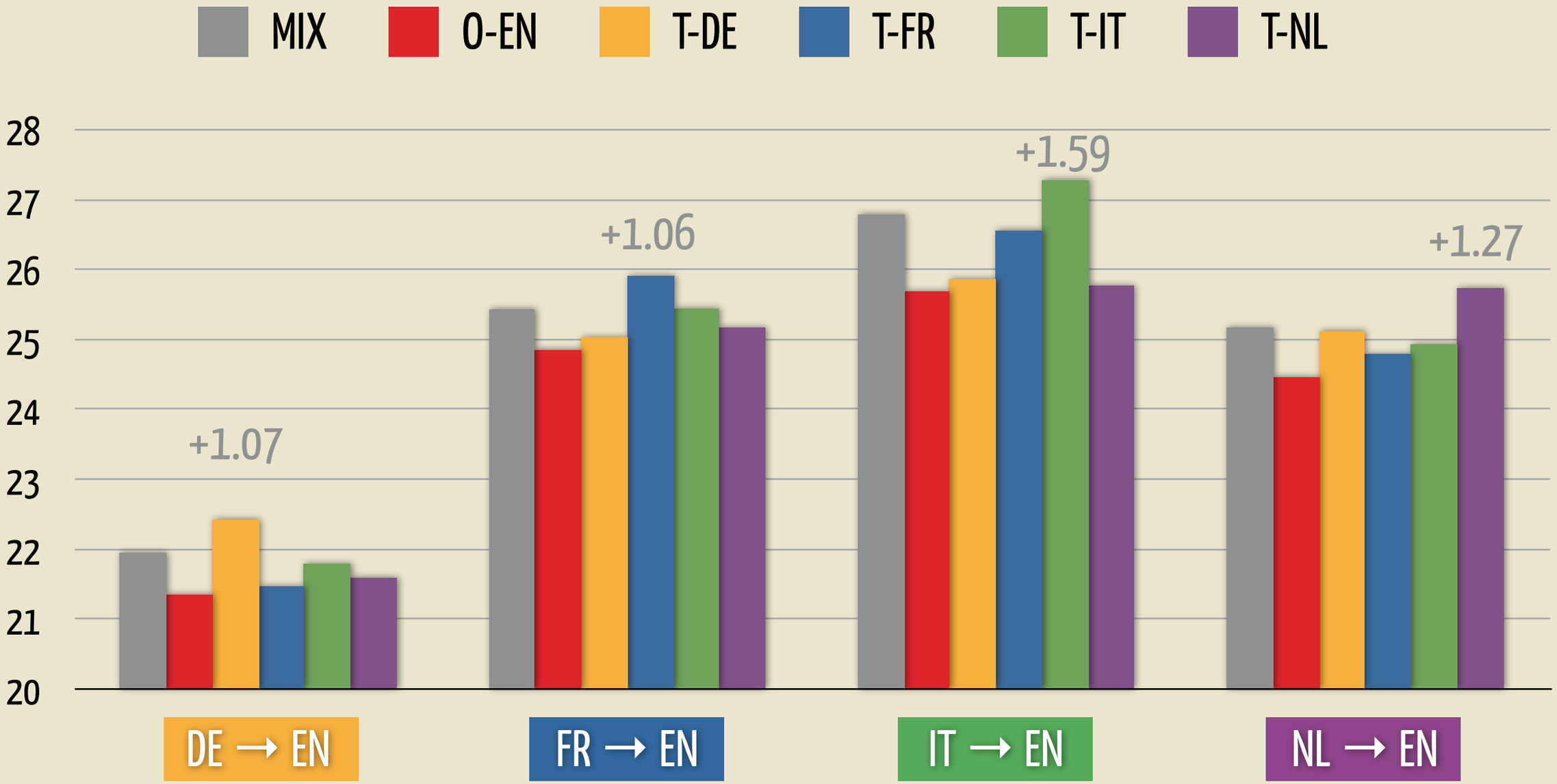


Perplexity - abstract

DE → EN

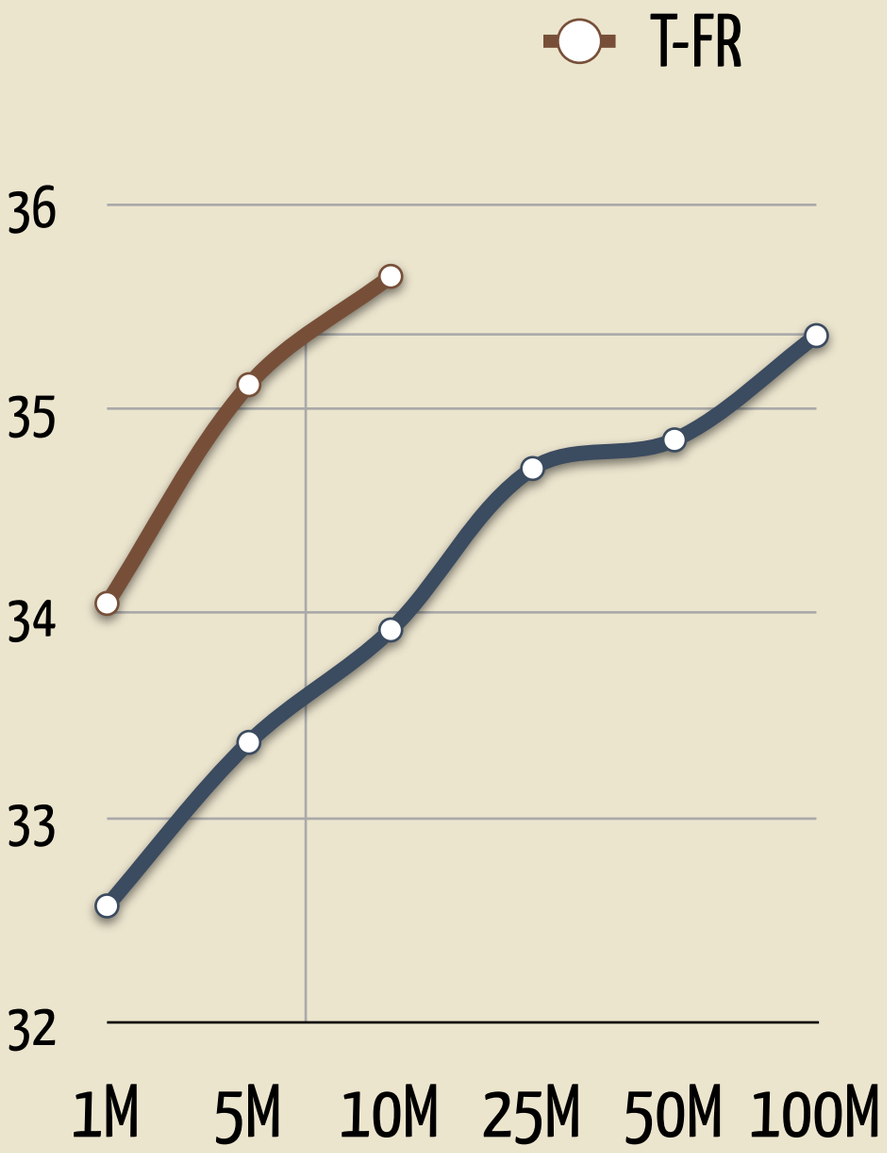


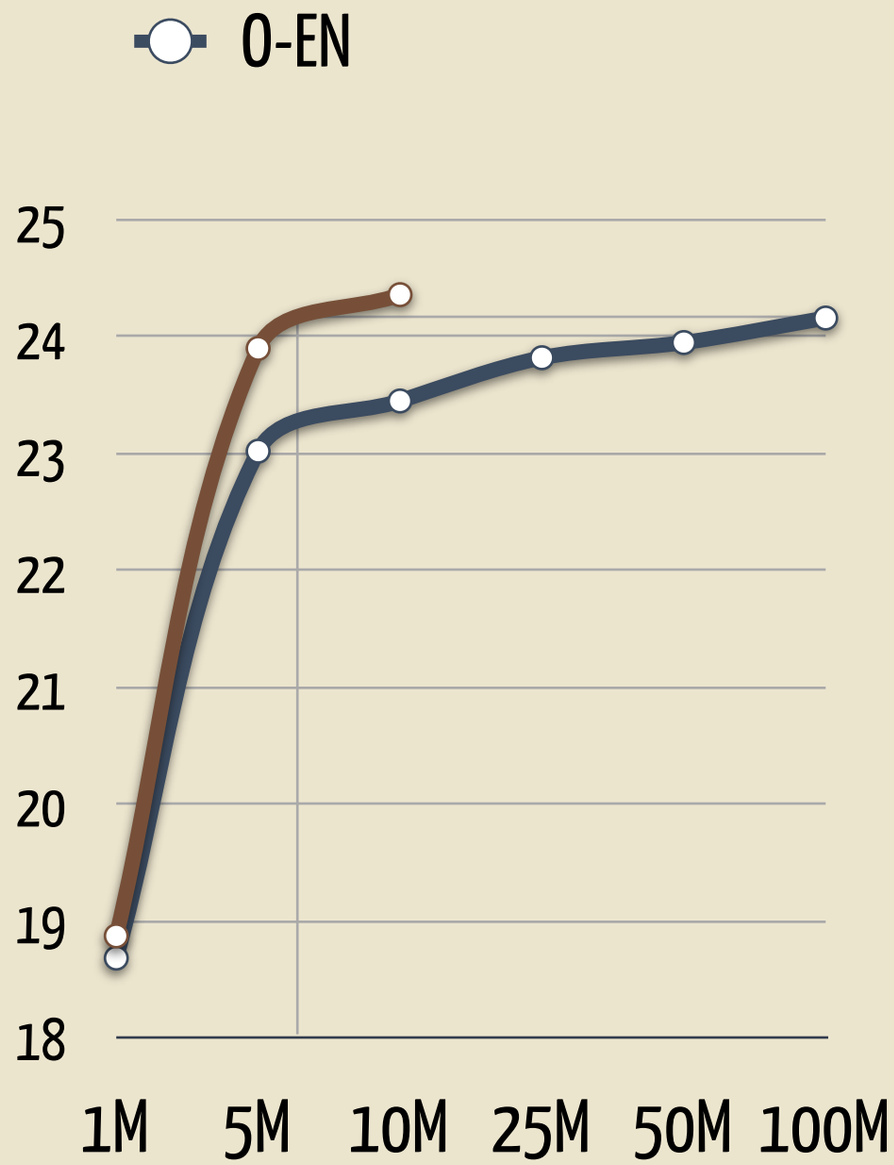
2. Machine Translation



3. LM Size / BLEU

FR → EN





Discussion

Is the number of tokens the appropriate unit?

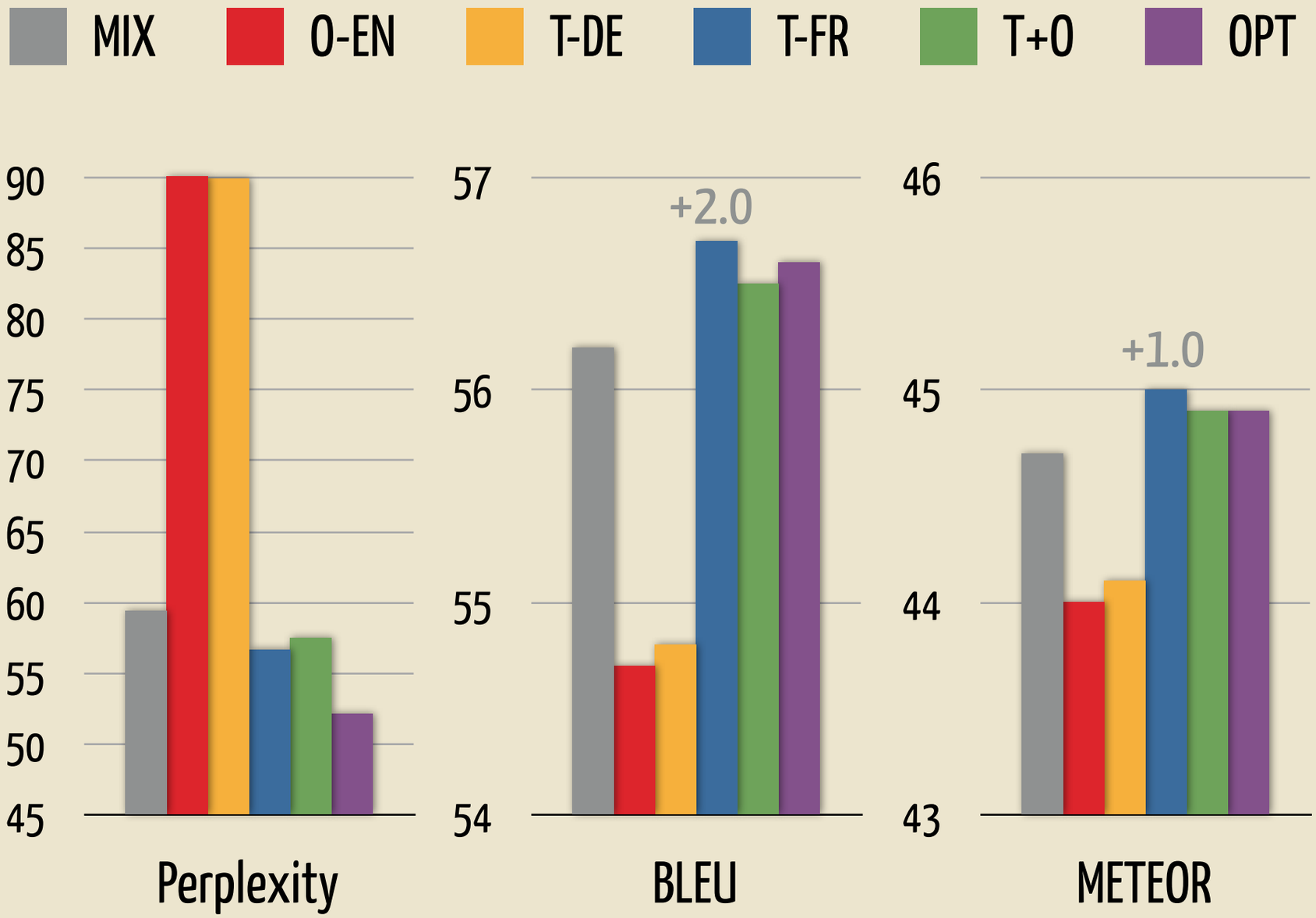
Do these claims hold:

- as more data is available?
- for other domains?

What to do when both original & translated available?

- combine
- discard original

Some additional experiments...



~ 70k docs / 1M sentences / 40M tokens

Conclusion

Not all texts were created equal!

This is often ignored in MT...